10. ročník Žilina 2. – 4. 7. 2018
Otvorený softvér vo vzdelávaní, výskume a v IT riešeniach
osscconf.soit.sk
OSS Conf

OSSConf 2024: 1–8

Novinky ze světa R+koronaviru

PAVEL STŘÍŽ (CZ)

Abstrakt. Článek je stručný popis myšlenek z doby koronaviru, kdy jsem pracoval s R. Pojďme trochu zavzpomínat na tu dobu.

Klíčová slova. R, CRAN, Bioconductor.

NEWS FROM THE WORLD OF R+CORONAVIRUS

Abstract. This article briefly describes my memories of the coronavirus period while working with R. Let us refresh the memories!

Keywords. R, CRAN, Bioconductor.

The R Foundation Retweeted: Peter Dalgaard. R 4.0.0 "Arbor Day" (source version) has been released.

24. 4. 2020 mi přistála na stole tato zpráva a za pár dní na to, 28. 4. 2020, došlo k aktualizaci Bioconductoru na verzi 3.11. Je Svátek práce, jdu ty nové `lazyLoad`, `lazyEval` a `lazyData` v R vyzkoušet a sdělit svůj nezávislý pohled.

1. R v4.0.0

Bez otálení jsem na svém Xubuntu 20.04 do `/etc/apt/sources.list` přidal:

```
deb https://cloud.r-project.org/bin/linux/ubuntu bionic-cran40/
```

Zakomentoval jsem starší pokusy a provedl preventivní kroky:

```
$ sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys
    E298A3A825C0D65DFD57CBB651716619E084DAB9
$ sudo apt update
$ sudo apt upgrade
```

Jádro aktualizace pak tvořil příkaz:

```
$ sudo apt install r-base r-base-dev
```

Rychlý test prokázal, že se podařilo a provedl jsem aktualizaci knihoven:

```
$ R --version
$ R
> update.packages(.libPaths())
```

2. COVID v19

The R Foundation Retweeted: R Consortium has started new GitHub repository to centralize collaboration and data sources – looking to develop COVID-19 tools and code – Come add your information and contribute to the community! *https://bddy.me/3aAXOmb*

Z toho samého dne zaujala mou pozornost ještě tato zpráva. Řekl jsem si, že bych měl nově nainstalované R hlouběji vyzkoušet.

Na úvodní stránce na mne vyskočily 4 projekty: Coronavirus Tracker, COVID-19 Propagation, COVID-19 Tracker Map a COVID-19 Projections.

Vezmu-li to od konce. U Projections na mne vyskočil GitHub. Po určitém bádání se mi podařilo otevřít rozcestník a webovou stránku pro Českou republiku, `https://www.volzinnovation.com/covid-19_SARS-CoV-2_corona/reports/latest/Czechia.html`. Autorem je Raphael Volz.

Autorem Tracker Map je Jay Ulfelder. V RStudiu Cloud mne hned upozornili, že se jedná o dočasný projekt. Pod Shiny app na mne vedle analýz vyškočily pdf v záložce WHO Situation Reports. Zajímavý nápad.

Druhý projekt v pořadí je Propagation. Autorem je Juan Francisco Venegas Gutiérrez. Bohužel repozitář na GitHubu mi nešel otevřít, tak jsem to zahlásil (Issues). Mezi modely jsem Českou republiku nenašel.

3. Coronavirus Tracker v0.1.4

První projekt v pořadí od Johna Coeneho Coronavirus Tracker vyzývá ke spuštění R. Pustil jsem se do toho. RStudio cloud občas jel bez přístupových práv, požadavek na knihovny `shinyMobile` a `echarts4r` zněl zajímavě.

```
$ R
> install.packages("coronavirus")
```

Zkusil jsem z dokumentace první ukázku a ještě si vyžádal knihovnu `dplyr`. Proč si ji nenainstaloval sám?

```
> install.packages("dplyr")
> library(coronavirus)
> require(dplyr)
> coronavirus %>% filter(type=="confirmed") %>% group_by(Country.Region) %>%
    summarise(total=sum(cases)) %>% arrange(-total) %>% head(20)
# A tibble: 20 x 2
   Country.Region       total
   <chr>                <int>
 1 Mainland China       70446
 2 Others                 355
 3 Singapore               75
```

Tady jsem zbystřil. To jsou stará data v textové formě, nikoliv hezké mapy přes web s aktuálními daty. Ve slangové řeči: rtfm! To, co jsem právě nainstaloval,

je knihovna `https://github.com/RamiKrispin/coronavirus`, která má stejný název. Mezi Issues jsem autorovi Trackeru zahlásil, že jeho název je v konfliktu s existující knihovnou z února 2020.

Pokračoval jsem v experimentování, knihovnu jsem odinstaloval a podíval se na návod, `https://coronavirus.john-coene.com`.

```
> remove.packages("coronavirus")
> install.packages("remotes")
> remotes::install_github("Johncoene/coronavirus")
```

Během instalace mi naskočila tato neobvyklá zpráva:

```
- Use `usethis::browse_github_pat()` to create a Personal Access Token.
- Use `usethis::edit_r_environ()` and add the token as `GITHUB_PAT`.
```

Knihovna `usethis` se teprve instalovala. V každém případě zmínka o `Rate limit reset at: [...]`, kdy došlo k restartu za několik minut pro mne znamenalo chvíli počkat a instalaci zopakovat.

Před koncem instalace si R vyžádalo systémový balík `libpq-dev` ve starší verzi 10.3-1 (aktuální je 10.12-0). Udělal jsem hrubý krok, doporučuji čtenářům najít lepší řešení přes Docker či pečlivě projít, co se bude odinstalovávat.

```
$ sudo apt install aptitude
$ sudo aptitude install libpq-dev
```

Na první dotaz jsem dal nikoliv (`n`; starší balík by se nenainstaloval), na druhou nabídku ano (`y`). Zopakoval jsem instalaci v R a skončila úspěšně.

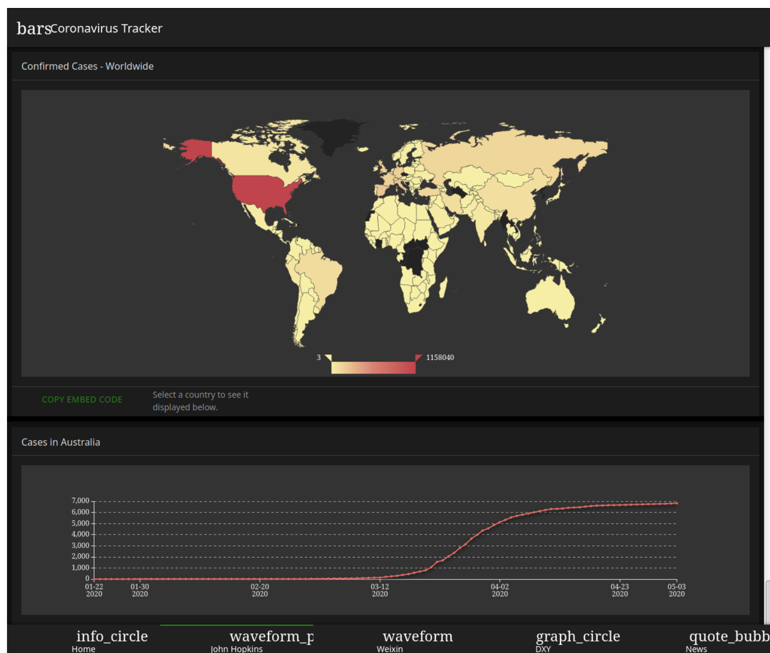
Zkusil jsem třířádkovou ukázku dle instalačního manuálu, `coronavirus.john-coene.com` (nutno zalistovat na webové stránce níž).

```
> library(coronavirus)
> virus<-crawl_coronavirus()
i Crawling data from John Hopkins
i Crawling data from Weixin
i Crawling data from DXY
> run_app(virus)
```

Pozn. Pokud bychom se dostali do konfliktu u příkazů, užijme:

```
> virus <- coronavirus::crawl_coronavirus()
> coronavirus::run_app(virus)
```

Ve webovém prohlížeči se mi otevřela vygenerovaná stránka, pokaždé na jiném portu. Radost byla veliká! Za pozornost stojí, že má být Johns Hopkins, to již někdo zahlásil autorovi k opravení.



Ve spodní části je pět záložek. Na druhé (waveform_path) v bloku *China* a *World* a čtvrté (graph_circle) v bloku *Cities* jako kdyby něco chybělo. Po nakliknutí do bloku se otevře detailní výpis. Vrátit zpět se dá přes značku xmark_circle_fill v pravém horním rohu. Design je trochu nezvyklý, ale musíme mít na paměti, že je to zaměřené na mobilní telefony a já to zkoušel na notebooku.

Detaily kolem dat je možné nalézt v levém horním rohu pod bars nebo menu. Z R lze server zavřít přes klávesovou kombinaci Ctrl+C.

Nyní dokumentace radí si nastavit `crontab` atd. Co mne zaujalo u stažení dat z DXY je, že se občas nezadařilo připojit. V rychlosti jsem nahlédl na server `https://education.rstudio.com/`, konkrétně na `dataio`.

Jakmile se podařilo na servery připojit, mohl jsem si proměnnou `virus` uložit a opětovně užívat. Rychlá pomůcka u experimentů bez nutnosti aktualizace dat.

```
> save(virus, file="virus.RData")
> load("virus.RData")
```

4. Hašovací klíč od `newsapi.org` v2

Mou pozornost zaujala poslední záložka se zprávou `No newsapi token`. To bych rád poléčil. Autor v dokumentaci radí:

```
> library(coronavirus)
> create_config()
```

V pozadí se ze šablony vytvoří soubor `_coronavirus.yml`, blok `database` je povinný, blok `newsapi` volitelný. To byl pro mne problém. Já jsem to chtěl obráceně. Nevadí.

Přes `https://newsapi.org/register` jsem se zaregistroval a získal hašovací klíč. Zahlédl jsem jejich novou knihovnu pro R `newsanchor`, my zůstaneme u autorem užité knihovny `newsapi`.

Zkusil jsem R podsunout hašovací klíč:

```
> library(newsapi)
> newsapi::newsapi_key("41e22e9efcf64b2a9354a796b99c43b8")
```

Ale ani touto cestou ani jinou přes editaci souboru `_coronavirus.yml` se mi to nepodařilo.

Prvně jsem nahlédl na zdrojové kódy v:

```
$ cd ~/R/x86_64-pc-linux-gnu-library/4.0/coronavirus
```

Narazil jsem hlavně na binární soubory rds, rdx a rdb. Nejsem expert, abych dokázal odpovědět, jestli by se soubory daly rozluštit a editovat.

Prozkoumal jsem zdrojové kódy přímo od autora:

```
$ git clone https://github.com/JohnCoene/coronavirus
```

Došel jsem k závěru, že bych musel zdrojové kódy upravit, zkompilovat atd. To je nad rámec této sváteční zprávy.

V souboru `coronavirus/inst/app/Dockerfile` jsem si ověřil, že skutečně knihovnu `newsapi` přebírá z GitHubu od uživatele `news-r`.

5. PostgreSQL v10+190

Říkal jsem si, když už se mi podařilo nainstalovat `libpq-dev`, dokáži i zbytek. Otevřel jsem komunitní tutoriál. Nainstaloval jsem PostgreSQL:

```
$ sudo apt install postgresql postgresql-contrib
```

A nejkratší možnou cestou jsem se pustil do dalších kroků. Vytvořil jsem v databázovém systému nového uživatele `testing` a novou databázi `testing`. Vynechávám krok vytvoření uživatele pod operačním systémem.

```
$ sudo -i -u postgres createuser --interactive
Enter name of role to add: testing
Shall the new role be a superuser? (y/n) y
$ sudo -u postgres createdb testing
```

Uživatel je bez hesla, to webové rozhraní nepřijme. Nastavil jsem nové heslo přes:

```
$ sudo -i -u postgres
$ psql
postgres=# ALTER USER testing WITH PASSWORD 'testing';
postgres=# \q
```

```
$ exit
```

Rychlokurz `psql`: `help` je základní nápověda, `\l` je výpis databází, `\c testing` je připojení k naší databázi, `\dt` je výpis tabulek, `\h` je seznam SQL příkazů, `\?` je seznam příkazů `psql` a `\q` ukončí běh programu. Verzálky u příkazů netřeba psát.

Vše zrealizované jsem zaznačil v `_coronavirus.yml`:

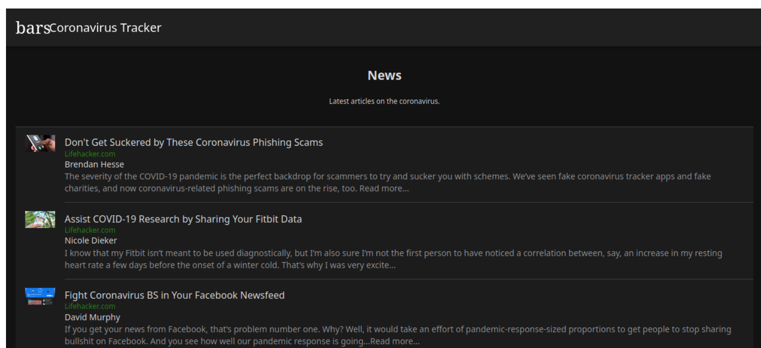
```
database:
  name: testing
  host: 127.0.0.1
  user: testing
  password: testing
newsapi:
  key: 41e22e9efcf64b2a9354a796b99c43b8
```

Když jsem opakoval tři řádky ukázkového spuštění v R, výpis se mi rozšířil o tyto dva řádky:

```
i Crawling news from newsapi.org
✔ Writing to database
```

V páté záložce mi vyběhly novinky, aktivovaný dotaz lze nalézt u autora v souboru `coronavirus/R/crawl.R`:

```
news <- newsapi::every_news("coronavirus OR covid", results = 100, language =
    "en", sort = "popularity")
```



Ověření funkčnosti můžeme zjistit i z tabulky `log`:

```
$ psql -h localhost -d testing -U testing
Password for user testing: testing
testing=# SELECT * FROM log;
```

Dostáváme přibližně takový výsledek:

```
         last_updated
-------------------------------
 2020-05-01 18:19:24.547171+02
(1 row)
```

Dle chuti lze dál bádat u surových dat, např.:

```
testing=# SELECT * FROM jhu WHERE country='Czechia';
testing=# SELECT * FROM jhu WHERE country='Slovakia';
testing=# \q
```

6. Bioconductor v3.11

Před dalším krokem si obvykle nastavuji plná práva u těchto adresářů:

```
$ cd /usr/lib/R
$ sudo chmod -R 777 site-library/
$ sudo chmod -R 777 library/
$ cd /usr/share
$ sudo chmod -R 777 R/
```

Za zmínku stojí, že manažer knihoven `biocLite` ustupuje a roli nahrazuje `BiocManager`. V R lze otestovat:

```
> install.packages("BiocManager")
> library(BiocManager)
> BiocManager::install()
> BiocManager::available()
```

Můžeme ověřit instalaci knihoven:

```
> BiocManager::valid()
   [...] "coronavirus", "echarts4r", "shinyMobile" [...]
Warning message:
0 packages out-of-date; 3 packages too new
```

To souhlasí, neb `coronavirus` byl instalován z GitHubu, nikoliv z CRANu.

Pro badatele stojí za pozornost obrazy pro Docker a Amazon (Amazon Machine Image, AMI). Je zde možnost instalovat vývojářské knihovny:

```
> BiocManager::install(version="devel")
```

7. Řešení konfliktu názvu knihovny v R

Na chvíli se ještě vraťme k řešení konfliktu stejného názvu knihoven. Na Stackoverflow zmiňují v principu dvě cesty.

Stáhnout si zdrojové soubory a nic neměnit:

```
$ cd /tmp
$ wget https://cran.r-project.org/src/contrib/coronavirus_0.1.0.tar.gz
$ R CMD INSTALL -l /tmp coronavirus_0.1.0.tar.gz
```

V R si pak volat jeden z příkazů a vybrat tak chtěnou knihovnu:

```
> # library(coronavirus)
> library(coronavirus, lib.loc="/tmp")
```

Když jsem zkoušel paralelně spustit i instalovanou knihovnu z GitHubu coronavirus odkomentováním prvního řádku, tak to neběželo. Tuším, že se jedná o bezpečnostní pojistku.

Druhá cesta je zasáhnout do souboru DESCRIPTION.

```
$ tar xvf coronavirus_0.1.0.tar.gz
$ mv coronavirus coronavirusRami
$ cd coronavirusRami
$ nano DESCRIPTION
```

První řádek upravit například na Package: coronavirusRami.

Volitelně upravíme i MD5, konkrétně první řádek za výpis:

```
$ md5sum -b DESCRIPTION
324f8275940bfa7fde376934c57a28ae *DESCRIPTION
```

Chtělo by to přejmenovat i další soubory na coronavirusRami, u této školní ukázky vynechávám. Zabalil jsem si zpět a už podsunul R:

```
$ cd ..
$ tar cvf coronavirusRami.tar.gz coronavirusRami/
$ R CMD INSTALL coronavirusRami.tar.gz
```

Ověřit funkčnost můžeme už přímo v R a lze si spustit oba konfliktní balíčky paralelně:

```
> library(coronavirus)
> ?coronavirus::crawl_coronavirus
> library(coronavirusRami)
> ?coronavirusRami::coronavirus
```

8. Pár tipů místo Závěru

Podobně jako jsou v R knihovny setříděné podle kategorií, viz Zobrazení úloh (CRAN Task Views), lze nahlédnout u RStudia na R Views s klíčovým slovem covid-19. K dnešnímu dni tam jsou tři záznamy: Some Select COVID-19 Modeling Resources, Simulating COVID-19 interventions with R a COVID-19 epidemiology with R.

Kdo by dal přednost odpočinku od R a PostgreSQL, nechť nahlédne na aktuální stavy kolem koronaviru na https://www.twitch.tv/killars.

Kontaktní adresa

Ing. Pavel Stříž, Ph.D., U Škol 940, Bučovice, okres Vyškov, 685 01, Česká republika,
E-mailová adresa: pavel@striz.cz